

A note on the instability and degeneracy of deep learning models

Andee Kaplan
Iowa State University
ajkaplan@iastate.edu and
Daniel Nordman
Iowa State University
dnordman@iastate.edu and
Stephen Vardeman
Iowa State University
vardeman@iastate.edu

Abstract

A probability model exhibits instability if small changes in a data outcome result in large, and often unanticipated, changes in probability. For correlated data structures found in several application areas, there is increasing interest in predicting/identifying instability. We consider the problem of quantifying instability for general probability models defined on sequences of observations, where each sequence of length N has a finite number of possible outcomes. (A sequence of probability models results indexed by N that accommodates data of expanding dimension.) Model instability is formally shown to occur when a certain log-probability ratio under such models grows faster than N . In this case, a one component change in the data sequence can shift probability by orders of magnitude. Also, as a measure of instability becomes more extreme, the resulting probability models are shown to tend to degeneracy, placing all their probability on arbitrarily small portions of the sample space. These results on instability apply to large classes of models commonly used in random graphs, network analysis, and machine learning contexts.

Keywords: Degeneracy, Instability, Deep Learning, Graphical Models

1 Introduction

We consider the behavior, and the potential impropriety, of probability models built to incorporate a sequence of discrete observations with length N . Let (X_1, \dots, X_N) denote a set of discrete random variables with a finite sample space, \mathcal{X}^N . That is, \mathcal{X} with $|\mathcal{X}| < \infty$ represents a finite set of potential outcomes for the variable X_i , and the data sequence (X_1, \dots, X_N) takes values in the N -fold product space \mathcal{X}^N . For each N , let P_{θ_N} denote a probability model on \mathcal{X}^N , under which $P_{\theta_N}(x_1, \dots, x_N) > 0$ is the probability of the data outcome $(x_1, \dots, x_N) \in \mathcal{X}^N$. We assume that the model support of P_{θ_N} is the sample space \mathcal{X}^N . This framework produces a series P_{θ_N} of probability models, indexed by a generic sequence of parameters θ_N , to describe data of each length $N \geq 1$. (The size and structure of such parameters are without restriction, and natural cases include those where $\theta_N \in \mathbb{R}^{q(N)}$ for some arbitrary integer-valued function $q(\cdot) \geq 1$.) We will call this model class *Finitely Supported Finite Sequence (FSFS) models*.

Section 2 provides several examples of FSFS models commonly used in graph/network analysis and machine learning (i.e., deep learning models). Section 3 establishes formal results regarding the propriety of FSFS models with regard to stability. A FSFS probability model sequence exhibits instability if small changes in the components of a data outcome (x_1, \dots, x_N) can result in large changes in probability $P_{\theta_N}(x_1, \dots, x_N)$. The concept of instability, introduced in the field of statistical physics by Ruelle (1999), was extended to include a notion of detection and quantification for certain exponential family models by Schweinberger (2011). For similar exponential models, particularly in connection to random graphs/networks, Handcock (2003) considered (mean-based) characterizations for so-called model degeneracy, whereby a probability model places all mass on a small subset of the sample space and produces undesirably low variability in model outcomes. As described by Schweinberger (2011), model instability and model degeneracy are related by viewing degeneracy as an extreme or limiting form of instability. The instability results of Schweinberger (2011) were developed for the case of discrete exponential family models. The main results here concern a general measure of model instability, appropriate across the whole FSFS model class. This can be used to identify when certain maximal probabilities

in FSFS models are too extreme relative to the length N . In this case, a one component change in the data sequence may shift probability by orders of magnitude, and FSFS models are rigorously shown to become degenerate as the measure of instability increases. Lastly, Section 4 emphasizes the implications of our model propriety results and proofs of the main results appear in the Appendix.

2 Examples

Many model families fall under the umbrella of FSFS models. For illustration, this section presents three specific examples of FSFS models, including models with deep architectures.

2.1 Discrete exponential family models

For $\mathbf{X} = (X_1, \dots, X_N)$ discrete random variables with sample space \mathcal{X}^N , $|\mathcal{X}| < \infty$, consider an exponential family model for \mathbf{X} with probability mass function of the form

$$p_{N,\boldsymbol{\lambda}}(\mathbf{x}) = \exp \left[\boldsymbol{\eta}^T(\boldsymbol{\lambda}) \mathbf{g}_N(\mathbf{x}) - \psi(\boldsymbol{\lambda}) \right], \quad \mathbf{x} \in \mathcal{X}^N,$$

for fixed positive integers k and L denoting the dimensions of the parameter, $\boldsymbol{\lambda} \in \Lambda \subset \mathbb{R}^k$ and natural parameter $\boldsymbol{\eta} : \mathbb{R}^k \mapsto \mathbb{R}^L$ spaces, $\mathbf{g}_N : \mathcal{X}^N \mapsto \mathbb{R}^L$ a vector of sufficient statistics,

$$\psi(\boldsymbol{\lambda}) = \log \sum_{\mathbf{x} \in \mathcal{X}^N} \exp \left[\boldsymbol{\eta}^T(\boldsymbol{\lambda}) \mathbf{g}_N(\mathbf{x}) \right], \quad \boldsymbol{\lambda} \in \Lambda,$$

the normalizing function, and $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^k : \psi(\boldsymbol{\lambda}) < \infty, k \leq q(N)\}$ is the parameter space.

Defining $P_{\boldsymbol{\theta}_N}(\mathbf{x}) \equiv p_{N,\boldsymbol{\lambda}_N}(\mathbf{x})$ with $\boldsymbol{\theta}_N = \boldsymbol{\lambda}_N$ to be a sequence of elements of $\Lambda \subset \mathbb{R}^k$ and noting that $P_{\boldsymbol{\theta}_N}(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^N$, these discrete exponential family models are special cases of the FSFS models. Such exponential models arise with spatial data on a lattice (Besag 1974), network data (Wasserman and Faust 1994; Handcock 2003), and even binomial sampling with N iid Bernoulli random variables. (Note that for random graphs or networks with, say, m nodes, one may wish to consider $N = \binom{m}{2}$ edges as binary (presence/absence) variables X_i . In this case, the length N of data sequence may naturally increase as a function of m .) For these exponential models, the dimension of the parameter

$\boldsymbol{\theta}_N$ is the same for each N as $\boldsymbol{\theta}_N$ lies in a parameter space of fixed Euclidean dimension k . This need not be true for other types of FSFS models considered in Sections 2.2-2.3. Schweinberger (2011) considered instability in such exponential models (e.g., for random graphs) for sequences of fixed parameters $\boldsymbol{\theta}_N = \boldsymbol{\lambda} \in \mathbb{R}^k$, $N \geq 1$, of non-varying dimension k .

2.2 Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is an undirected graphical model specified for discrete or continuous random variables, binary variables being most common (cf. Smolensky 1986). A RBM architecture has two layers, hidden (\mathcal{H}) and visible (\mathcal{V}), with conditional independence within each layer. Let $\mathbf{X} = (X_1, \dots, X_N)$ denote the N random variables for visibles with support \mathcal{X}^N and $\mathbf{H} = (H_1, \dots, H_{N_H})$ denote the N_H random variables for hiddens with support \mathcal{X}^{N_H} where $\mathcal{X} = \{-1, 1\}$. For parameters $\boldsymbol{\alpha} \in \mathbb{R}^{N_H}$, $\boldsymbol{\beta} \in \mathbb{R}^N$, and Γ as a matrix with dimension $N_H \times N$, the RBM for $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{H})$ then has the joint probability mass function

$$P_{\boldsymbol{\theta}_N}(\tilde{\mathbf{x}}) = \exp \left[\boldsymbol{\alpha}^T \mathbf{h} + \boldsymbol{\beta}^T \mathbf{x} + \mathbf{h}^T \Gamma \mathbf{x} - \psi(\boldsymbol{\theta}_N) \right], \quad \tilde{\mathbf{x}} = (\mathbf{h}, \mathbf{x}) \in \mathcal{X}^{N+N_H},$$

where

$$\psi(\boldsymbol{\theta}_N) = \log \sum_{\tilde{\mathbf{x}} \in \mathcal{X}^{N+N_H}} \exp \left[\boldsymbol{\alpha}^T \mathbf{h} + \boldsymbol{\beta}^T \mathbf{x} + \mathbf{h}^T \Gamma \mathbf{x} \right], \quad \boldsymbol{\theta}_N \in \Theta_N,$$

is the normalizing function. Let $\boldsymbol{\theta}_N = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Gamma) \in \Theta_N \subset \mathbb{R}^{q(N)}$ with $q(N) = N + N_H + N * N_H$ denote the vector of parameters for the RBM.

The probability mass function for the visible variables X_1, \dots, X_N follows from marginalizing this joint specification:

$$P_{\boldsymbol{\theta}_N}(\mathbf{x}) = \sum_{\mathbf{h} \in \mathcal{X}^{N_H}} P_{\boldsymbol{\theta}_N}(\mathbf{x}, \mathbf{h}), \quad \mathbf{x} \in \mathcal{X}^N.$$

Note that the vector of model parameters $\boldsymbol{\theta}_N$, of size $q(N)$, grows in size as a function of sample dimension N to accommodate the dimension of visible variables X_1, \dots, X_N , and one may further choose the number N_H of hidden variables to change with N as well; in

particular, the number N_H of hidden units may also potentially increase with N . Additionally, as $|\mathcal{X}| = 2$ and $P_{\theta_N}(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^N$, the RBM model specification for visibles is a FSFS model. This example also indicates that models formed by marginalizing a base FSFS model (e.g., a type of exponential family model) is again a FSFS model class.

2.3 Deep learning

Consider two models with “deep architecture” that contain multiple hidden (or latent) layers in addition to a visible layer of data, a deep Boltzmann machine (Salakhutdinov and Hinton 2009) and a deep belief network (Hinton, Osindero, and Teh 2006). Let M denote the number of hidden layers included in the model and $N_{(H,1)}, \dots, N_{(H,M)}$ the number of hidden variables within each hidden layer. Then let $\tilde{\mathbf{X}} = \{H_1^{(1)}, \dots, H_{N_{(H,1)}}^{(1)}, \dots, H_1^{(M)}, \dots, H_{N_{(H,M)}}^{(M)}, \mathbf{X}\}$ be random variables corresponding to the hidden variables $\{H_i^{(j)} : i = 1, \dots, N_{(H,j)}, j = 1, \dots, M\}$ and visible variables $\mathbf{X} = (X_1, \dots, X_N)$ in a deep probabilistic model. Each variable outcome will again lie in $\mathcal{X} = \{-1, 1\}$.

Deep Boltzmann machine (DBM). The DBM class of models maintains conditional independence within all layers in the model by stacking RBM models and maintaining the restriction of only allowing conditional dependence between neighboring layers. The joint probability mass function for a DBM is

$$P_{\theta_N}(\tilde{\mathbf{x}}) = \exp \left[\sum_{i=1}^M \boldsymbol{\alpha}^{(i)T} \mathbf{h}^{(i)} + \boldsymbol{\beta}^T \mathbf{x} + \mathbf{h}^{(1)T} \Gamma^{(0)} \mathbf{x} + \sum_{i=1}^{M-1} \mathbf{h}^{(i)T} \Gamma^{(i)} \mathbf{h}^{(i+1)} - \psi(\boldsymbol{\theta}_N) \right],$$

for $\tilde{\mathbf{x}} = (\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(M)}, \mathbf{x}) \in \mathcal{X}^{N_{(H,1)} + \dots + N_{(H,M)} + N}$ where

$$\psi(\boldsymbol{\theta}_N) = \log \sum_{\tilde{\mathbf{x}} \in \mathcal{X}^{N_{(H,1)} + \dots + N_{(H,M)} + N}} \exp \left[\sum_{i=1}^M \boldsymbol{\alpha}^{(i)T} \mathbf{h}^{(i)} + \boldsymbol{\beta}^T \mathbf{x} + \mathbf{h}^{(1)T} \Gamma^{(0)} \mathbf{x} + \sum_{i=1}^{M-1} \mathbf{h}^{(i)T} \Gamma^{(i)} \mathbf{h}^{(i+1)} \right],$$

for $\boldsymbol{\theta}_N \in \Theta_N$ is the normalizing function and parameters in the model are $\boldsymbol{\beta} \in \mathbb{R}^N$, $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^{N_{(H,i)}}$ for $i = 1, \dots, M$, along with a matrix $\Gamma^{(0)}$ of dimension $N_{(H,1)} \times N$, and matrices $\Gamma^{(i)}$ of dimension $N_{(H,i)} \times N_{(H,i+1)}$ for $i = 1, \dots, M-1$. Let $\boldsymbol{\theta}_N = (\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(M)}, \boldsymbol{\beta}, \Gamma^{(0)}, \dots, \Gamma^{(M-1)}) \in \Theta_N \subset \mathbb{R}^{q(N)}$ denote the combined vector of parameters with total length $q(N) = N_{(H,1)} + \dots + N_{(H,M)} + N + N_{(H,1)} * N + N_{H,2} * N_{(H,1)} + \dots + N_{(H,M)} * N_{(H,M)-1}$.

The probability mass function for the visible random variables X_1, \dots, X_N follows from this joint specification as

$$P_{\theta_N}(\mathbf{x}) = \sum_{(\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(M)}) \in \mathcal{X}^{N_{(H,1)} + \dots + N_{(H,M)}}} P_{\theta_N}(\tilde{\mathbf{x}}), \quad \mathbf{x} \in \mathcal{X}^N$$

Again like the RBM case, the visible DBM model specification is an example of a FSFS model.

Deep belief network (DBN). A DBN resembles a DBM in that there are multiple layers of latent random variables stacked in a deep architecture with no conditional dependence between layers. The difference between the DBM and DBN models is that all but the last stacked layer in a DBN are Bayesian networks (see Pearl 1985), rather than RBMs. Thus for visibles X_1, \dots, X_N with support \mathcal{X}^N , a DBN is also a FSFS model if the number $q(N)$ of components in the parameter vector is dependent on the dimension of the visibles. Commonly, as in logistic belief nets (Neal 1992), a “weight” parameter is placed on each interaction between visibles, X_1, \dots, X_N and the first layer of latent variables, $H_1^{(1)}, \dots, H_{N_{(H,1)}}^{(1)}$, satisfying the definition of a FSFS model.

3 Instability results

To define or measure instability in FSFS models, it is useful to consider the behavior of a data model sequence P_{θ_N} . A relevant quantity to this end is a (scaled) extremal log-probability ratio (ELPR)

$$\frac{1}{N} \log \left[\frac{\max_{(x_1, \dots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \dots, x_N)}{\min_{(x_1, \dots, x_N) \in \mathcal{X}^N} P_{\theta_N}(x_1, \dots, x_N)} \right] \equiv \frac{1}{N} \text{ELPR}_N(\theta_N). \quad (1)$$

The main idea is that, in formulating FSFS models for potentially increasing numbers of variables (i.e., for $N \rightarrow \infty$), the ratio (1) should remain bounded, requiring that the largest probability possible under P_{θ_N} should maintain a fixed order of magnitude relative to the smallest probability allowed under the same model. Specifically, the log of the ratio should grow at mostly linearly with the sample size N . This leads to the following definition.

Definition 1 (S-unstable FSFS). Let $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$ be a sequence of FSFS model parameters where the size of the model $q(N)$ is a function of the number of random variables N . A FSFS model formulation is *Schweinberger-unstable* or *S-unstable* if, as the number of variables increase ($N \rightarrow \infty$),

$$\lim_{N \rightarrow \infty} \frac{1}{N} \text{ELPR}(\boldsymbol{\theta}_N) \equiv \lim_{N \rightarrow \infty} \frac{1}{N} \log \left[\frac{\max_{(x_1, \dots, x_N) \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(x_1, \dots, x_N)}{\min_{(x_1, \dots, x_N) \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(x_1, \dots, x_N)} \right] = \infty.$$

In other words, given any $C > 0$, there exists an integer $N_C > 0$ so that $\frac{1}{N} \text{ELPR}_N(\boldsymbol{\theta}_N) > C$ for all $N \geq N_C$.

This definition of *S-unstable* is a generalization or reinterpretation of “unstable” used in Schweinberger (2011) by allowing non-exponential family models (e.g. RBM and DBM models in Sections 2.2-2.3) and an increasing number of parameters. While this definition differs in form and scope from the original, it does match that in Schweinberger (2011) for the special case of exponential models (cf. Section 2.1) considered there.

S-unstable FSFS model sequences are undesirable for several reasons. One is that small changes in data can lead to overly-sensitive changes in probability. Consider, for example,

$$\Delta(\boldsymbol{\theta}_N) \equiv \max \left\{ \log \frac{P_{\boldsymbol{\theta}_N}(\mathbf{x})}{P_{\boldsymbol{\theta}_N}(\mathbf{x}^*)} : \mathbf{x} \text{ \& } \mathbf{x}^* \in \mathcal{X}^N \text{ differ in exactly one component} \right\},$$

the biggest log-probability ratio for a one component change in data outcomes at a FSFS parameter $\boldsymbol{\theta}_N$. We then have the following (non-asymptotic) result.

Proposition 1. *Let $\text{ELPR}(\boldsymbol{\theta}_N)$ be as in (1) for an integer $N \geq 1$. For a given $C > 0$, if*

$$\frac{1}{N} \text{ELPR}_N(\boldsymbol{\theta}_N) > C,$$

then

$$\Delta_N(\boldsymbol{\theta}_N) > C.$$

Again, if the probability ratio (1) is too large, then the FSFS model will exhibit large changes in probability for very small differences in the data configuration, which exemplifies the intuitive notation of instability.

Additionally, S-unstable FSFS model sequences are connected to degenerate models, where model *degeneracy* implies placing all probability on a small portion of the sample space. For perspective, note that differing bounds on $1/N \cdot \text{ELPR}(\boldsymbol{\theta}_N)$ in (1) provide a spectrum of levels of “stability” and Proposition 1 indicates increasing and undesirable sensitivity of model probabilities (i.e., for one component changes in outcomes) as (1) increases. Furthermore, as the instability measure (1) grows, FSFS model sequences are guaranteed to slide into full degeneracy as Proposition 2 will show. Define a ϵ -modal set

$$M_{\epsilon, \boldsymbol{\theta}_N} \equiv \left\{ \mathbf{x} \in \mathcal{X}^N : \log P_{\boldsymbol{\theta}_N}(\mathbf{x}) > (1 - \epsilon) \max_{\mathbf{x}^* \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}^*) + \epsilon \min_{\mathbf{x}^* \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}^*) \right\}$$

of possible outcomes, for a given $0 < \epsilon < 1$.

Proposition 2. *For an unstable FSFS model in Definition 1, and for any given $0 < \epsilon < 1$,*

$$P_{\boldsymbol{\theta}_N}((x_1, \dots, x_N) \in M_{\epsilon, \boldsymbol{\theta}_N}) \rightarrow 1 \text{ as } N \rightarrow \infty.$$

In other words, in S-unstable FSFS models, all probability in the model formulation with a large number of random variables will concentrate mass on mode sets of arbitrarily small size or simply on those few outcomes with the highest probability.

Remark 1. There is a further generalization the notion of instability in Definition 1 meant to address independent replications of data sequences. That is, one might consider data as n independent and identically distributed replications $\mathbf{X}_1, \dots, \mathbf{X}_n$, where each $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,N}) \in \mathcal{X}^N$ follows a common FSFS model with probabilities $P_{\boldsymbol{\theta}_N}(\mathbf{x}) > 0$, $\mathbf{x} \in \mathcal{X}^N$ and $|\mathcal{X}| < \infty$, for $i = 1, \dots, n$. This leads to a total of $n * N$ random variables in the joint model. However, the definition of S-unstable and Propositions 1-2 still hold for such iid replications. This is because $\left(\max_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}) \right)^n$ is the largest probability possible under the joint model for the n replications while $\left(\min_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}) \right)^n$ is the smallest probability. Thus, for the combined replications $\mathbf{X}_1, \dots, \mathbf{X}_n$, the analog definition of the extremal log-probability becomes

$$\begin{aligned} \frac{\text{extremal log-probability ratio}}{\# \text{ random variables in the model}} &\equiv \frac{1}{n * N} \log \left[\frac{\left(\max_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}) \right)^n}{\left(\min_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x}) \right)^n} \right] = \frac{1}{N} \log \left[\frac{\max_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x})}{\min_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x})} \right] \\ &= \frac{1}{N} \text{ELPR}(\boldsymbol{\theta}_N), \end{aligned}$$

implying that the definition of an S-unstable FSFS model sequence is invariant to the level (n) of independent replication. Consequently, overall model instabilities may be characterized by those of one observation from the common FSFS model.

4 Implications

For a large class of models that covers a broad range of applications (including “deep learning”), we have developed a formal definition and elucidated multiple consequences of instability. We have shown for FSFS models that instability manifests through small changes in data leading to overly-sensitive changes in probability as well as the potential to place all probability on a small piece of the sample space. Models that fall within the definition of a FSFS model should be used with caution to ensure that the effects of instability are not experienced.

A Appendix: Proofs of main results

Proof of Proposition 1. We prove the contrapositive, supposing that $\Delta(\boldsymbol{\theta}_N) \leq C$ holds for some $C > 0$ and show $\text{ELPR}(\boldsymbol{\theta}_N) \leq NC$. Let $\mathbf{x}_{min} \equiv \arg \min_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x})$ and $\mathbf{x}_{max} \equiv \arg \max_{\mathbf{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\mathbf{x})$. Note there exists a sequence $\mathbf{x}_{min} \equiv \mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_k \equiv \mathbf{x}_{max}$ in \mathcal{X}^N of component-wise switches to move from \mathbf{x}_{min} to \mathbf{x}_{max} in the sample space (i.e. $\mathbf{x}_i, \mathbf{x}_{i+1} \in \mathcal{X}^N$ differ in exactly 1 component for $i = 0, \dots, k$) for some integer $k \in \{0, 1, \dots, N\}$. Under the FSFS model, recall $P_{\boldsymbol{\theta}_N}(\mathbf{x}) > 0$ holds so that $\log P_{\boldsymbol{\theta}_N}(\mathbf{x})$ is well-defined for each outcome $\mathbf{x} \in \mathcal{X}^N$. Then, if $k > 0$, it follows that

$$\begin{aligned} \text{ELPR}(\boldsymbol{\theta}_N) &= \log \left[\frac{P_{\boldsymbol{\theta}_N}(\mathbf{x}_{max})}{P_{\boldsymbol{\theta}_N}(\mathbf{x}_{min})} \right] = \left| \sum_{i=1}^k \log \left(\frac{P_{\boldsymbol{\theta}_N}(\mathbf{x}_i)}{P_{\boldsymbol{\theta}_N}(\mathbf{x}_{i-1})} \right) \right| \\ &\leq \sum_{i=1}^k \left| \log \left(\frac{P_{\boldsymbol{\theta}_N}(\mathbf{x}_i)}{P_{\boldsymbol{\theta}_N}(\mathbf{x}_{i-1})} \right) \right| \leq k \Delta_N(\boldsymbol{\theta}_N) \leq NC, \end{aligned}$$

using $k \leq N$ and $\Delta(\boldsymbol{\theta}_N) \leq C$. If $k = 0$, then $\mathbf{x}_{max} = \mathbf{x}_{min}$ and the same bound above holds. \square

Proof of Proposition 2. Define \mathbf{x}_{max} and $\mathbf{x}_{min} \in \mathcal{X}^N$ as in the proof of Proposition 1 where $|\mathcal{X}| < \infty$ holds in the FSFS model. We may suppose $|\mathcal{X}| > 1$ (i.e., \mathcal{X}^N has more than one outcome) because otherwise the model is trivially degenerate for all $N \geq 1$. Fix $0 < \epsilon < 1$. Then, $\mathbf{x}_{max} \in M_{\epsilon, \theta_N}$, so $P_{\theta_N}(M_{\epsilon, \theta_N}) \geq P_{\theta_N}(\mathbf{x}_{max}) > 0$. If $\mathbf{x} \in \mathcal{X}^N \setminus M_{\epsilon, \theta_N}$, then by definition $P_{\theta_N}(\mathbf{x}) \leq [P_{\theta_N}(\mathbf{x}_{max})]^{1-\epsilon} [P_{\theta_N}(\mathbf{x}_{min})]^\epsilon$ holds so that

$$\begin{aligned} 1 - P_{\theta_N}(M_{\epsilon, \theta_N}) &= \sum_{\mathbf{x} \in \mathcal{X}^N \setminus M_{\epsilon, \theta_N}} P_{\theta_N}(\mathbf{x}) \\ &\leq (|\mathcal{X}|^N) [P_{\theta_N}(\mathbf{x}_{max})]^{1-\epsilon} [P_{\theta_N}(\mathbf{x}_{min})]^\epsilon. \end{aligned}$$

From the lower bound on $P_{\theta_N}(M_{\epsilon, \theta_N})$ and the upper bound on $1 - P_{\theta_N}(M_{\epsilon, \theta_N})$, it follows that

$$\begin{aligned} \frac{1}{N} \log \left[\frac{P_{\theta_N}(M_{\epsilon, \theta_N})}{1 - P_{\theta_N}(M_{\epsilon, \theta_N})} \right] &\geq \frac{1}{N} \log \left[\frac{P_{\theta_N}(\mathbf{x}_{max})}{(|\mathcal{X}|^N) [P_{\theta_N}(\mathbf{x}_{max})]^{1-\epsilon} [P_{\theta_N}(\mathbf{x}_{min})]^\epsilon} \right] \\ &= \frac{\epsilon}{N} \log \left[\frac{P_{\theta_N}(\mathbf{x}_{max})}{P_{\theta_N}(\mathbf{x}_{min})} \right] - \log |\mathcal{X}| \rightarrow \infty \end{aligned}$$

as $N \rightarrow \infty$ by the definition of an unstable FSFS model (c.f. Definition 1). Consequently, $P_{\theta_N}(M_{\epsilon, \theta_N}) \rightarrow 1$ as $N \rightarrow \infty$ as claimed. \square

References

- Besag, Julian. 1974. “Spatial Interaction and the Statistical Analysis of Lattice Systems.” *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 192–236.
- Handcock, Mark S. 2003. “Assessing Degeneracy in Statistical Models of Social Networks.” Center for Statistics; the Social Sciences, University of Washington. <http://www.csss.washington.edu/>.
- Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. 2006. “A Fast Learning Algorithm for Deep Belief Nets.” *Neural Computation* 18 (7). MIT Press: 1527–54.
- Neal, Radford M. 1992. “Connectionist Learning of Belief Networks.” *Artificial Intelligence* 56 (1). Elsevier: 71–113.
- Pearl, Judea. 1985. “Bayesian Networks: A Model of Self-Activated Memory for Evidential

Reasoning.” UCLA Computer Science Department.

Ruelle, D. 1999. *Statistical Mechanics: Rigorous Results*. London: Imperial College Press.

Salakhutdinov, Ruslan, and Geoffrey E Hinton. 2009. “Deep Boltzmann Machines.” In *International Conference on Artificial Intelligence and Statistics*, 448–55. AI & Statistics.

Schweinberger, Michael. 2011. “Instability, Sensitivity, and Degeneracy of Discrete Exponential Families.” *Journal of the American Statistical Association* 106 (496). Taylor & Francis: 1361–70.

Smolensky, Paul. 1986. “Information Processing in Dynamical Systems: Foundations of Harmony Theory.” DTIC Document.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge: Cambridge University Press.